San Diego State University

Antoni Luque Santolaria <aluque@sdsu.edu>

More "super-short" Caudovirales data

Stephen Nayfach <snayfach@lbl.gov> Wed, Jan 13, 2021 at 12:30 PM
To: Antoni Luque Santolaria <aluque@sdsu.edu>

Hi Antoni,

> Got it, Stephen. Thanks for the quick response. I'll let you and Simon judge if a map would be compelling when you get a chance to evaluate the draft that I'm preparing. If needed, the coordinates associated with the contigs that came from the Global Ocean Virome should be accessible. I assume that combining the IMG and GOV would be a good chunk of original data.

Sounds like a good plan.

> Here is another request. Let's see if this one is more doable. Do you have easy access to a variable, for example, genome content (in bases), to compare the contribution of each biome-level (1, 2, 3) to the complete dataset where you extracted the circular genomes (<20 kb)?

Unfortunately, not. But I do have access to the full set of circular genomes (>2kb) for each biome, which I could use to calculate the cumulative genome length per biome. Would that be helpful?

> I want to test if the frequency of circular contigs (<20 kb) observed across biomes is similar to the proportion of the genetic content that each biome contributed to the total dataset. The null hypothesis is that both distributions would be similar. The assumption behind the hypothesis is that all biomes would have a similar relative content of small phages. Significant differences between the distributions of small circular contigs vs. the genetic contribution of a biome to the total dataset would indicate that some biomes are richer (or poorer) in circular small phages.

We can certainly test the ratio of small vs large Caudovirales phages using the information on hand.

Also, thinking back to a previous comment you made, all the sequences I sent did have terminal repeats (I believe I clipped the end off in the final dataset) and should certainly be viral. So we may want to go back and include some of those that have been excluded. But, I'm also happy to proceed with the current set you have for the initial analysis, as we can always go back later to add more sequences. For example, I've now assembled close to 4,000 nanopore datasets, which could be really useful to give us more confidence these small genomes are bona fide.

> Total genomic content: 30,000 Mb (Megabases)
>
> Biome level 1:
> Host-associated --> 25,000 Mb
> Aquatic --> 10,000 Mb
> ...
>
> Let me know if this makes any sense to you and if it is possible.

Please see attached for numbers of total genome length of circular Caudovirales from each biome.

Best,
Stephen

3 attachments

biome2.tsv
2K

biome3.tsv
5K

biome1.tsv
1K